Sliced Wasserstein Variational Inference

Mingxuan Yi MINGXUAN.YI@BRISTOL.AC.UK and **Song Liu** SONG.LIU@BRISTOL.AC.UK
School of Mathematics, University of Bristol, UK

Editors: Emtiyaz Khan and Mehmet Gönen

Abstract

Variational Inference approximates an unnormalized distribution via the minimization of *Kullback-Leibler* (KL) divergence. Although this divergence is efficient for computation and has been widely used in applications, it suffers from some unreasonable properties. For example, it is not a proper metric, i.e., it is non-symmetric and does not preserve the triangle inequality. On the other hand, optimal transport distances recently have shown some advantages over KL divergence. With the help of these advantages, we propose a new variational inference method by minimizing sliced Wasserstein distance–a valid metric arising from optimal transport. This sliced Wasserstein distance can be approximated simply by running MCMC but without solving any optimization problem. Our approximation also does not require a tractable density function of variational distributions so that approximating families can be amortized by generators like neural networks. Furthermore, we provide an analysis of the theoretical properties of our method. Experiments on synthetic and real data are illustrated to show the performance of the proposed method.

Keywords: Variational inference, Sliced Wasserstein distance, Optimal transport, Markov chain Monte Carlo

1. Introduction

Variational inference (VI) is a method that recasts Bayesian inference as an optimization problem where it uses *Kullback-Leibler* (KL) divergence as a measurement to capture the discrepancy of two probability distributions. Unlike the traditional inference methods that utilize Monte Carlo Markov Chains (MCMC) to sample from the target probability space, VI is fast and lightweight in terms of computation. Therefore, it is preferred in many modern machine learning tasks.

Optimal Transport (OT) (Villani, 2009) has recently gained significant attentions in the machine learning community. Compared to KL divergence, OT gives a valid metric that is symmetric and preserves triangular inequality. It is reported to show good performances in some downstream applications (Arjovsky et al., 2017; Gulrajani et al., 2017). While OT provides us with a new horizon on some old machine learning scenarios, the original OT problem requires a computationally demanding optimization procedure which impedes the popularity of applying the original methods. To address this difficulty, sliced Wasserstein distance (Bonnotte, 2013; Bonneel et al., 2015) reduces the computational inefficiency of OT by projecting high dimensional probability distributions into univariate slices where OT problem has a closed-form solution. Similar to the standard Wasserstein distance, sliced Wasserstein distance is still a valid metric function (Bonnotte, 2013). This metric has been

successfully used in many practical tasks (Deshpande et al., 2018; Kolouri et al., 2018b,a) but it has not yet been studied in variational inference. See Figure 1 for the fitted Gaussian distributions (dotted and solid lines) obtained by minimizing sliced Wasserstein distance and various information divergences. It is known that VI tends to search modes with the reverse KL divergence but tends to spread the mass with the forward KL divergence. Sliced Wasserstein distance has a different behavior compared to three other information divergences in Figure 1. In Figure 1 (b), the approximating distribution is initialized to the right side. The sliced Wasserstein distance produces a closer fit to the higher mode than all other information divergences. Sliced Wasserstein distance trades off mode-seeking and mass spreading and the approximating distribution is less sensitive to model misspecification.

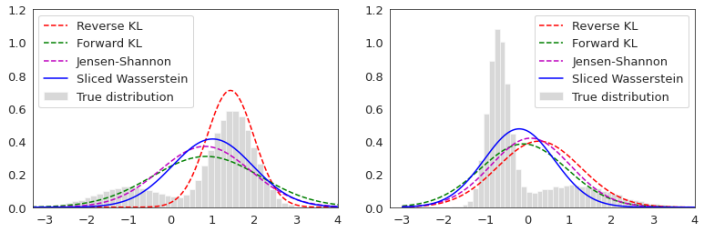


Figure 1: The visualization of performances of using different discrepancies to match two probability distributions. *Jensen-Shannon* divergence is a symmetric information divergence via averaging reverse and forward KL divergences.

In this paper, we extend sliced Wasserstein distance to variational inference tasks. The distance between the variational and the target distribution is approximated using MCMC. Consequently, we sequentially minimize sliced Wasserstein distance between the variational distribution and the marginal distributions. As the MCMC evolves, such a procedure guides the variational distribution to match the target under the minimization of a metric function. One advantage is that by leveraging the sliced Wasserstein distance, our method does not rely on simultaneous adversarial training (Mescheder et al., 2017; Li et al., 2017; Zhang et al., 2020) to estimate the discrepancy but can still perform amortized inference (Gershman and Goodman, 2014), i.e., use a parameterized function as a black-box sampler, e.g. neural networks, to capture target distributions.

Contributions: i). We develop a new variational inference method that minimizes sliced Wasserstein distance – a statistical distance. ii). The convergence and asymptotic properties of the proposed method are discussed as well as burn-in analysis. iii) We provide empirical studies to illustrate the method via experiments.

2. Background

2.1. Variational Inference

Given an unnormalized probability distribution $\bar{\mu}(z)$, it could be difficult to obtain the normalizing constant $\beta = \int \bar{\mu}(z)dz$ and as such, we would not have assess to the true density $\mu(z) = \beta^{-1}\bar{\mu}(z)$. Variational inference aims to find a distribution $\upsilon_\phi(z)$ which approximates $\mu(z)$ as close as possible. Such approximations can be obtained via minimizing *Kullback-*

Leibler (KL) divergence

$$D_{KL}\big[\upsilon_\phi||\mu\big] = \int \upsilon_\phi(z) \log \frac{\upsilon_\phi(z)}{\mu(z)} dz \tag{1}$$

Note that $D_{KL}\big[\upsilon_\phi||\mu\big] = 0$ if and only if $\upsilon_\phi(z) = \mu(z)$. However, Eq(1) is intractable because the density function $\mu(z)$ is known up to a normalizing constant. Instead, we can equivalently maximize *Evidence Lower Bound* (ELBO)

$$\log \beta \geq \mathcal{L}(\phi) = \mathbb{E}_{\upsilon_\phi(z)}\big[\log \bar{\mu}(z) - \log \upsilon_\phi(z)\big] \tag{2}$$

Since we observe that the model evidence $\log \beta$ is a constant w.r.t. ϕ and the above inequality becomes tight if $D_{KL}\big(\upsilon_\phi||\mu\big) = 0$. Optimization of ELBO requires the differentiation of the r.h.s. expectation. Gradient descent is a standard approach that allows for such optimization. To obtain a valid estimation of the gradient, a solution is to apply the score function method (Paisley et al., 2012; Ranganath et al., 2014). An alternative solution to obtain the gradient of ELBO is the reparameterization trick (Kingma and Welling, 2014; Rezende et al., 2014). Vanilla VI leverages KL divergence but this can be substituted with any other f-divergences and importance sampling (Jerfel et al., 2021; Wan et al., 2020; Prangle, 2019) can be used to obtain gradient estimation for general f-divergences.

2.2. Wasserstein Distance

Wasserstein distance arises in optimal transport (Villani, 2009) in which a distribution is transformed to another by moving probability mass. Wasserstein distance measures the cost of such a transformation. We denote $\mathcal{X}$ the sample space and let $\mathcal{Q}_p(\mathcal{X})$ be the set of Borel probability measures with finite p-th moment. Given two marginal distributions $\mu(x), \upsilon(y) \in \mathcal{Q}_p(\mathcal{X})$, let $\Pi(\mu, \upsilon)$ be a set of any coupled joint distributions $\gamma(x, y)$ where $\int_{\mathcal{X}} \gamma(x, y)dx = \upsilon(y)$ and $\int_{\mathcal{X}} \gamma(x, y)dy = \mu(x)$. The $p-$Wasserstein distance is defined as

$$\mathcal{W}_p(\mu, \upsilon) = \left\{ \inf_{\gamma \in \Pi(\mu,\upsilon)} \int_{\mathcal{X}\times\mathcal{X}} \|x - y\|^p d\gamma(x, y) \right\}^{\frac{1}{p}} \tag{3}$$

where $\|x - y\|$ is a cost function of moving a mass from μ to υ. Intuitively, the $p-$Wasserstein distance aims to find an optimal joint distribution $\gamma(x, y)$ where the expected cost specified by Eq(3) achieves its minimum. Solving this optimization problem is generally difficult (Cuturi, 2013), but we can rewrite $p-$Wasserstein distance in a univariate case as

$$\mathcal{W}_p(\mu, \upsilon) = \left\{ \int_0^1 \left| F_\mu^{-1}(t) - F_\upsilon^{-1}(t) \right|^p dt \right\}^{\frac{1}{p}} = \left\{ \int_{\mathcal{X}} \left| x - F_\upsilon^{-1}(F_\mu(x)) \right|^p d\mu(x) \right\}^{\frac{1}{p}} \tag{4}$$

where $F(\cdot)$ is a cumulative distribution function and $F^{-1}(\cdot)$ is a quantile function of a probability distribution and the composition $F_v^{-1}F_u(\cdot)$ defines a transportation map that moves mass from $u(x)$ to $\upsilon(y)$. Given two empirical distributions, we can simply utilize Eq(4) to estimate $p-$Wasserstein distance by approximating the quantile function with sorted samples.

2.3. Sliced Wasserstein Distance

Motivated by the computational efficiency of estimating Wasserstein distance with univariate distributions. We give a brief review of sliced Wasserstein distance (Bonneel et al., 2015). We first introduce *Radon* transformation (Beylkin, 1984).

Let $h(\cdot)$ be a function $h : \mathbb{R}^d \longrightarrow \mathbb{R}$. The *Radon* transform is defined as

$$\mathcal{R}h_\theta(l) = \int_{S:l=\langle x,\theta\rangle} h(x)dS \tag{5}$$

Eq(5) defines a surface integral on a hyper-plane $S : l = \langle x, \theta\rangle$ where $l \in \mathbb{R}$ and $\theta \in \mathbb{S}^{d-1}$ where $\mathbb{S}^{d-1}$ is a unit ball embedded in $\mathbb{R}^d$. For any pair of vectors θ and h, we obtain a sliced function $h^R_\theta(\cdot)$. We note that marginalization of a high dimensional joint probability distribution can be regarded as a special case of the *Radon* transform with $\theta = e_i$, where e_i is an all-zero vector with only 1 at the i-th position. Note that the sliced function yielded by Eq(5) is univariate. Leveraging this property, we define sliced Wasserstein distance for probability distributions $\mu(x)$ and $\upsilon(y)$ as the average distance resulting from these slices.

$$\mathcal{SW}_p(\mu, \upsilon) = \left(\int_{\theta\in\mathbb{S}^{d-1}} \mathcal{W}_p^p(\mathcal{R}\mu_\theta, \mathcal{R}\upsilon_\theta)d\theta\right)^{\frac{1}{p}} \tag{6}$$

Given an empirical distribution described by $\mu^n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, it is trivial to write down its *Radon* transformation defined in Eq(5) as $\mathcal{R}\mu^n_\theta = \frac{1}{n}\sum_{i=1}^n \delta_{\langle x_i,\theta\rangle}$. We summarize the procedure of calculating sliced Wasserstein distance via empirical samples in **Algorithm 1**. The sample sorting in step 2 corresponds to calculating quantile functions in the univariate Wasserstein distance Eq(4).

Algorithm 1 Estimation of Sliced Wasserstein Distance with Samples

Require: $\mu^n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\upsilon^n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$ (for simiplicity, we assume two distributions have the same number of observations)
for $k = 0, 1 \cdots m$

1. Sample θ_k from $\mathbb{S}^{d-1}$ uniformly,
2. Obtain slices and sort $\{\langle x_i, \theta_k\rangle\} \longrightarrow \{\langle x_j, \theta_k\rangle\}$ and $\{\langle y_i, \theta_k\rangle\} \longrightarrow \{\langle y_j, \theta_k\rangle\}$

return $\mathcal{SW}_p(\mu^n, \upsilon^n) = \left(\frac{1}{mn}\sum_{k=1}^m\sum_{j=1}^n \left|\langle x_j, \theta_k\rangle - \langle y_j, \theta_k\rangle\right|^p\right)^{\frac{1}{p}}$

3. Sliced Wasserstein Variational Inference

We name our method as sliced Wasserstein variational inference (SWVI). We use the following notation: $\upsilon_\phi(z)$ as the variational distribution parameterized by ϕ, and $\mu(z) = \beta^{-1}\bar{\mu}(z)$ as the target distribution, where β is a normalizing constant. The problem we are interested

in is finding an optimal parameter ϕ^* that minimizes sliced Wasserstein distance between the variational distribution and the target distribution.

$$\phi^* = \arg\min_\phi \mathcal{SW}_p(\mu, \upsilon_\phi) \tag{7}$$

Eq(7) defines a minimum distance estimator (Wolfowitz, 1957) (Basu et al., 2011) where we choose sliced Wasserstein distance as a specific metric function. In some applications, the density function of υ_ϕ is not always tractable but we can still simulate samples from it and if μ also has a sampling distribution, we solve the following problem,

$$\phi^* = \arg\min_\phi \mathcal{SW}_p(\mu^n, \upsilon_\phi^n) \tag{8}$$

where μ^n and υ_ϕ^n are sampling distributions of μ and υ_ϕ with n observations. The optimization problem defined in Eq(8) does not require an explicit probability density function– allowing to design a more flexible variational distribution such as a neural network generator and a variational program (Ranganath et al., 2016).

3.1. Estimation of Sliced Wasserstein Distance

Due to the intractability of the target distribution $\mu(z) = \beta^{-1}\bar{\mu}(z)$, the main idea of SWVI is to approximate sliced Wasserstein distance between the variational and the target distribution using MCMC and minimize it. Unlike variational inference, MCMC methods provide particle approximations by designing transition kernels of a Markov chain with invariant distribution $\mu(z)$. Let $K(\cdot|\cdot)$ be a transition kernel of MCMC, and $\mu_0(z)$ be the initial distribution of the corresponding MCMC, e.g., a potential $\mu_0(z)$ can be chosen as $\upsilon_\phi(z)$. We denote by $\mu_t(z)$ the marginal distribution of the Markov chain after applying t times transitions.

$$\mu_t(z) = \int \mu_{t-1}(z')K(z|z')dz' \tag{9}$$

Given a sufficiently long run, $\mu_t(z)$ converges to $\mu(z)$ because of the stationary property of Markov chain. At the current stage, one can directly evaluate sliced Wasserstein distance $\mathcal{SW}_p(\mu, \upsilon_\phi)$ via

$$\mathcal{SW}_p(\mu, \upsilon_\phi) = \mathcal{SW}_p(\mu_t, \upsilon_\phi) \text{ as } t \to \infty \tag{10}$$

Unfortunately, running a long enough MCMC chain is time consuming and it might be difficult to diagnose the burn-in period. To solve this problem, we instead evaluate a local distance $\mathcal{SW}_p(\mu_t, \upsilon_\phi)$ at every time step t of iterating MCMC algorithms. Next we update parameters ϕ via gradient descent to minimize $\mathcal{SW}_p(\mu_t, \upsilon_\phi)$ per iteration. Since every $\mu_t(z)$ is an improvement of the previous $\mu_{t-1}(z)$, minimizing sliced Wasserstein distance guides the variational distribution $\upsilon_\phi(z)$ towards the target distribution $\mu(z)$.

Note that we use particle approximations to the marginal distribution μ_t by parallelizing n Markov chains and we sample the same number of particles from the variational distribution υ_ϕ. Hence, the optimization problem is replaced with minimizing the sampled-based sliced Wasserstein distance $\mathcal{SW}_p(\mu_t^n, \upsilon_\phi^n)$. Therefore, SWVI defines the following sequential optimization problem

$$\{\min_\phi \mathcal{SW}_p(\mu_t^n, \upsilon_\phi^n)\}_{t\in\mathbb{N}} \tag{11}$$

Heuristically, we also inherit the parameter from the previous iteration such that $\phi_t = \arg\min_\phi \mathcal{SW}_p(\mu_t^n, \upsilon_{\phi_{t-1}}^n)$. We summarize this procedure in **Algorithm 2**.

Algorithm 2 Sliced Wasserstein Variational Inference (SWVI)

Require: An unnormalized probability distribution $\bar{\mu}(z)$, a Markovian transition kernel $K(z|z')$ with invariant distribution $\mu(z)$ and initial distribution $\mu_0(z)$, variational distribution $\upsilon_\phi(z)$.
Initialize $\phi_0 = \phi$ and $\mu_0^n(z)$ by sampling n particles from $\mu_0(z)$
for $t = 1, 2 \cdots T$

1. Apply transition kernel $K(z|z')$ to μ_{t-1}^n to get μ_t^n
2. $\phi_t = \arg\min_\phi \mathcal{SW}_p(\mu_t^n, \upsilon_{\phi_{t-1}}^n)$

return $\upsilon_{\phi_T}(z)$

3.2. Existence, Convergence and Consistency of SWVI

It is important to understand the convergence of SWVI while the algorithm iterating with time t and the asymptotic property if the number of Markov chains n goes to infinity. Specifically, we show that SWVI converges if the Markov chain converges to the invariant distribution (**Theorem 1**) and SWVI is consistent (**Theorem 2**) if the number of Markov chains goes to infinity under some mild regularities. Inspired by (Bernton et al., 2019) (Nadjahi et al., 2019), we make the following assuamptions

Assumption A.1. For any sequence of probability measures $\{\mu_t\}_{t\in\mathbb{N}}$, e.g., marginal probability densities yield by Markov chains, $\{\mu_t\}_{t\in\mathbb{N}}$ converges in sliced Wasserstein distance to μ, i.e. $\lim_{t\to\infty} \mathcal{SW}_p(\mu_t, \mu) = 0$, $\mathbb{P}$-almost surely.
Assumption A.2. The map $\phi \to \upsilon_\phi$ is continuous, i.e., $\lim_{k\to\infty} ||\phi_k - \phi|| = 0$ implies weak convergence of υ_{ϕ_k} to υ_ϕ.
Assumption A.3. For any data generating processes $\{\mu_t^n\}_{n\in\mathbb{N}}$, we have $\lim_{n\to\infty} \mathcal{SW}_p(\mu_t^n, \mu_t) = 0$. For $\lim_{n\to\infty} ||\phi_n - \phi|| = 0$, we have $\lim_{n\to\infty} \mathcal{SW}_p(\upsilon_{\phi_n}^n, \upsilon_\phi) = 0$, $\mathbb{P}$-almost surely.
Assumption A.4. For some $\epsilon > 0$ with $\epsilon^* = \inf_\phi \mathcal{SW}_p(\mu_t, \upsilon_\phi)$, $B_\epsilon = \{\phi \in \mathcal{H} : \mathcal{SW}_p(\mu_t, \upsilon_\phi) \leq \epsilon^* + \epsilon\}$ is bounded.

Note that all infimum are taken with $\phi \in \mathcal{H}$ where $\mathcal{H}$ is a parameteric space. **A.1** indicates the convergence of Markov chain in the metric space. Indeed, it also implies weak convergence $\{\mu_t\}_{t\in\mathbb{N}} \to \mu$ in $\mathcal{Q}_p(\mathcal{X})$. This is a corollary from (Nadjahi et al., 2020) where if p-Wasserstein distance metrizes weak convergence then the sliced p-Wasserstein metrizes weak convergence as well. We refer to (Villani, 2009) for the study of weak convergence with Wasserstein distance. A straightforward result is that Langevin dynamic is a gradient flow of KL divergence in Wasserstein Space (Jordan et al., 1998) and (Liu et al., 2019) generalizes it to other MCMC algorithms.

Theorem 1 *Under* ***Assumption A.1*** *and* ***A.2****,*

$$\lim_{t\to\infty} \inf_{\phi} \mathcal{SW}_p(\mu_t, \upsilon_\phi) = \inf_{\phi} \mathcal{SW}_p(\mu, \upsilon_\phi) \tag{12}$$

Theorem 2 *Under* ***Assumption A.2****,* ***A.3*** *and* ***A.4***

$$\lim_{n\to\infty} \inf_{\phi} \mathcal{SW}_p(\mu_t^n, \upsilon_\phi^n) = \inf_{\phi} \mathcal{SW}_p(\mu_t, \upsilon_\phi) \tag{13}$$

and the existence of minimum

$$\arg\min_{\phi} \mathcal{SW}_p(\mu_t, \upsilon_\phi) \neq \emptyset$$

Corollary 3 *If Markov chains evolve with time and the number of chains goes to infinity, we have*

$$\lim_{t\to\infty} \lim_{n\to\infty} \inf_{\phi} \mathcal{SW}_p(\mu_t^n, \upsilon_\phi^n) = \inf_{\phi} \mathcal{SW}_p(\mu, \upsilon_\phi) \tag{14}$$

Theorem 1 indicates that the sequential optimization problem of SWVI converges. The proof is analogous to (Nadjahi et al., 2019) where the sequence of probability measures here is constructed by Markov chains. **Theorem 2** is similar to Corollary 6.11 by Villani (2009) where if μ^n and υ^n weakly converge in $\mathcal{Q}_p(\mathcal{X})$ to μ and υ, then $\lim_{n\to\infty} \mathcal{W}_p(\mu^n, \upsilon^n) = \mathcal{W}_p(\mu, \upsilon)$. **Corollary 3** is a direct result from **Theorem 1** and **2**. Here we prove the infimum of the sequence $\{\mathcal{SW}_p(\mu_t^n, \upsilon_\phi^n)\}_{n\in\mathbb{N}}$ converges to the infimum of $\mathcal{SW}_p(\mu_t, \upsilon_\phi)$ in Appendix A.

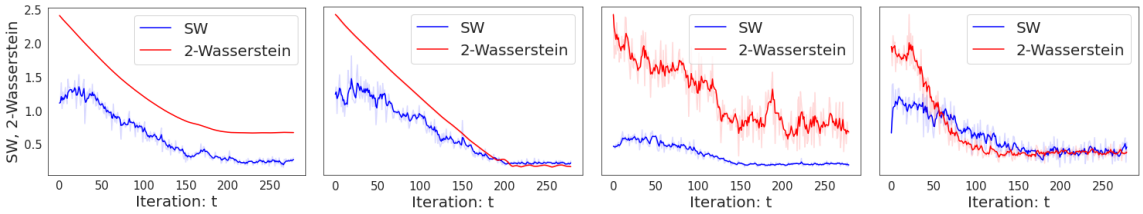
3.3. Burn-in Diagnosis

SWVI utilizes MCMC to explore the target space. Hence the convergence of the corresponding Markov chain influences the accuracy of the variational approximation. Burn-in diagnosis of MCMC is a hard problem in the machine learning and statistics community. Several attempts are proposed to assess the convergence such as estimating marginal densities and using couplings to estimate an upper bound of Wasserstein distance (Biswas et al., 2019).

We show that our method does not suffer from the problem of determining the burn-in period of Markov Chain. Sliced Wasserstein distance itself can monitor the convergence of MCMC. Without of the loss of generality, we assume that the burn-in period $t < M$ for some $M \in \mathbb{N}$ such that,

$$\mathcal{SW}_p(\mu_{t+1}, \upsilon_\phi) \approx \mathcal{SW}_p(\mu_t, \upsilon_\phi) \approx \mathcal{SW}_p(\mu, \upsilon_\phi), \text{ for } t \geq M, \text{for all } \phi \tag{15}$$

This implies that $\min_\phi \mathcal{SW}_p(\mu_t, \upsilon_\phi)$ becomes stationary for $t \geq M$. That is to say once the loss function achieves the minimum with a tolerance, SWVI congerves. We show the loss curves of our method of fitting 2D Gaussian distributions (corresponds to experiment **4.1**) in Figure 2 and compare it with 2-Wasserstein distance where it is tractable (we use sinkhorn divergence Cuturi (2013) as approxiamted 2-Wasserstein distance for mixture distributions).



(a) (b) (c) (d)

Figure 2: Losses of fitting Gaussian distributions. (a) Mean-field Gaussian approximations. (b)-(d) Full Gaussian approximations where (c) and (d) has a target bi-modal Gaussian mixtures. Compared with 2-Wasserstein (Sinkhorn approximations in (c) and (d)) distances between Gaussians. SWVI curves have empirically similar convergence rates to the convergence described by 2-Wasserstein distances

3.4. Stochastic Optimization

The sampled-based sliced Wasserstein distance $\mathcal{SW}_p(\mu_t^n, \upsilon_\phi^n)$ can be estimated by drawing samples respectively from $\mu_t(z)$ (via MCMC) and $\upsilon_\phi(z)$. Suppose that $\{z_i\}_{i=1,2\cdots n} \sim \upsilon_\phi(z)$ and $\{z_i'\}_{i=1,2\cdots n} \sim \mu_t(z)$. Sliced Wasserstein distance is then approximated by

$$\mathcal{SW}_p(\mu_t^n, \upsilon_\phi^n) \approx \mathcal{L}(\{z_i\}, \{z_i'\}) \tag{16}$$

Here we rewrite the approximate distance as a function $\mathcal{L}(\cdot, \cdot)$ of two sets of samples according to **Algorithm 1**. In order to optimize the parameter of the variational distribution $\upsilon_\phi(z)$, we still need to reparameterize samples $\{z_i\}_{i=1,2\cdots n}$. This can be done by an amortized sampler that can be either a parametric probability distribution or a flexible neural network generator. The amortized sampler is written as $z(\phi) = g_\phi(\epsilon),\ \epsilon \sim r(\epsilon)$, where $r(\epsilon)$ is a noise distribution and g_ϕ is a parametric model. We can use chain rule to obtain the gradient estimation of Eq(16),

$$\nabla_\phi \mathcal{L}(\{z_i\}, \{z_i'\}) = \sum_{i=1}^{n} \nabla_{z_i} \mathcal{L}(\{z_i\}, \{z_i'\}) \nabla_\phi z_i(\phi). \tag{17}$$

This can be implemented easily via back-propagation. We also find it is helpful to warm-up MCMC by a few steps and then applies SWVI iteration. The only difference of such lagged procedure is that SWVI starts at a better initial distribution of MCMC. The practical implementation of SWVI is summarized in **Algorithm 3**

3.5. Related Work

To the best of our knowledge, we first introduce sliced Wasserstein distance into variational inference. Sliced Wasserstein has been widely studied in generative modeling (Deshpande et al., 2018; Kolouri et al., 2018a; Bonet et al., 2021) as a measurement of the discrepancy of model distributions and data. Ambrogioni et al. (2018) combines orginal Wasserstein distances with variational inference where they introduce a new class of discrepancies that includes f-divergences allowing for variational inference problems. On the other hand, many

Algorithm 3 Practical Implementation of SWVI

Require: An unnormalized probability distribution $\bar{\mu}(z)$, a Markovian transition kernel $K(z|z')$ with invariant distribution $\mu(z)$ and initial distribution $\mu_0(z)$, variational distribution $\upsilon_\phi(z)$, learning rate α and warm-up lag L.
Initialize $\phi_0 = \phi$ and $\mu_0^n(z)$ by sampling n particles from $\mu_0(z)$
for $t = 1, 2, \cdots L-1$

1. Only run MCMC, i.e., apply kernel $K(z|z')$ to μ_{t-1}^n to get μ_t^n

for $t = L, L+1 \cdots T$

1. Apply transition kernel $K(z|z')$ to μ_{t-1}^n to get μ_t^n with corresponding particles $\{z_i'\}_{i=1,2\cdots n} \sim \mu_t(z)$
2. Draw $\{z_i\}_{i=1,2\cdots n} \sim \upsilon_{\phi_t}(z)$ with reparameterization
3. Update parameter $\phi_{t+1} = \phi_t - \alpha \nabla_\phi \mathcal{L}(\{z_i\}, \{z_i'\})$

return $\upsilon_{\phi_T}(z)$

alternative objectives to (reverse) KL divergence has been proposed in VI problems. For example, Stein discrepancy (Ranganath et al., 2016; Liu and Wang, 2016), forward KL divergence (Jerfel et al., 2021; Prangle, 2019; Bornschein and Bengio, 2014; Dieng et al., 2017), α- divergence (Li and Turner, 2016) and f-divergences (Wan et al., 2020). Since our methods uses MCMC to estimate the metric function. This is similar to some previous works utilize the advantages of MCMC methods to improve variational inference. The authors of (Ruiz and Titsias, 2019) proposes to use MCMC samples to estimate a new objective function to ELBO. The authors of (Naesseth et al., 2020) estimates the gradient of forward KL divergence via runing MCMC. Similar to our work, The authors of (Li et al., 2017) uses 'teacher-student' framework where MCMC samples teach the variational distribution to how to improve via minimizing different objectives. Another line of research to enrich the approximating families of VI is normalizing flows (Rezende and Mohamed, 2015). Compared to our method, normalizing flows require manually design a bijective function whereas SWVI can use a simple non invertible neural net.

4. Experiments

For all experiments, we use sliced 1-Wasserstein distance. Details of experiment settings can be found in Appendix B.

4.1. Toy Experiment

In this experiment, we set target distributions as a 2D Gaussian distribution and a bi-modal Gaussian mixture. We fit the variational distribution to the target distribution via vanilla variational inference under reverse KL divergence and the proposed method SWVI. In our method, we adopt the random walk Metropolis-Hastings algorithm as our MCMC instance.

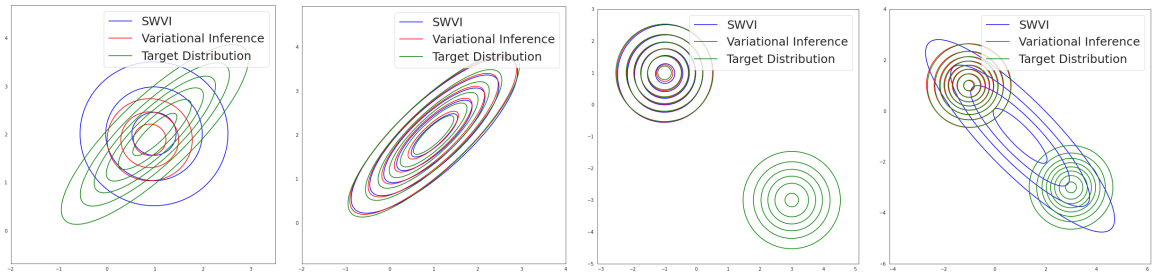


Figure 3: Approximating 2D synthetic distributions with VI and SWVI

In 1st figure of Figure 3, we use a mean-field Gaussian distribution as the variational distribution and it shows that SWVI results in an approximation with a larger variance compared to standard VI. In the 2nd figure, we use a regular Gaussian distribution with fully trainable covariance matrix, both VI and SWVI can approximate the target distribution well. For the 3rd and 4th figures, the target distribution is set as a Gaussian mixture model, VI always fits one mode but SWVI can have different behaviours, i.e., jumping between different local minima, if we choose different step sizes in MCMC (random walk with standard deviations 0.2 and 2.5). Note that Wasserstein-2 distance only has closed-form expressions between two Gaussians so minimising Wasserstein-2 distance is not tractable on the third and fourth examples of Figure 3. In the first example, the approximating distribution of Wasserstein-2 distance is a Gaussian with the same mean and a diagonal covariance whose diagonal elements are the same as the target distribution.

4.2. Implicit Variational Distribution with Neural Nets

We show an experiment where we have amortized SWVI such that the variational distribution is a neural network generator without an explicit density function. Note that the proposed method in **Algorithm 3** does not require a closed form of density function of $\upsilon_\phi(z)$. Hence, we can easily adapt a neural net that can generate samples from a more flexible distribution. For comparison, we also implemented amortized Stein variational gradient (SVGD) method (Liu and Wang, 2016; Feng et al., 2017). The target distribution is a mixture of two Gaussians but one with a larger variance and another with a smaller variance. We fit a neural net generator to this distribution with SWVI and amortized SVGD. This experiment shows that amortized SVGD fails to capture the other mode. The reason is that the kernel function (RBF) cannot adjust the bandwidth to the two modes with different ranges of variance. However, with the asymptotic guarantees of MCMC, the generator trained with SWVI can efficiently capture two different modes and outputs considerably better samples.

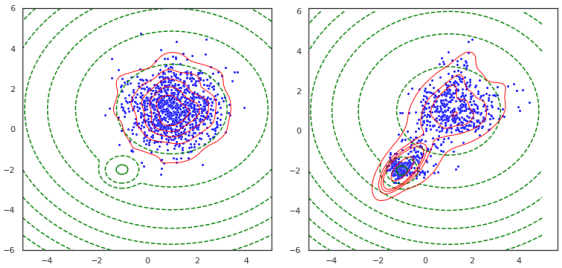


Figure 4: Generated samples (blue dots) with SVGD (L) and SWVI (R).

4.3. Bayesian Logistic Regression

We apply SWVI to binary classification tasks in the UCI repository (Asuncion and Newman, 2007) with Bayesian logistic regression models. We set the prior distribution to be a constant for the Bayesian logistic model and Langevin Dynamics (with learning rate 0.0001) (Neal, 2011; Welling and Teh, 2011) as the MCMC instance. The variational distributions are both mean-field Gaussians for SWVI and VI. We present results in Table 1 and it can be seen that the performance of SWVI is on par with the vanilla VI.

Dataset	Mean-field SWVI	Mean-field VI
Heart	0.852±0.019	**0.855**±0.030
Wine	0.716±0.025	**0.731**±0.012
Ionosphere	**0.771**±0.071	0.767±0.062

Table 1: Test accuracy for Bayesian logistic regression (32 posterior samples are used). The dimensions of parameters are 11, 34 and 13 respectively.

4.4. Latent Generative Modeling

We consider a latent generative model where observations x is supposed to be generated by a latent variable z. In this model, θ parameterizes a likelihood function $p_\theta(x|z)$. Given a prior $p(z)$ for the latent variable, we have a posterior distribution

$$\mu_\theta(z|x) \propto p(z)p_\theta(x|z) \tag{18}$$

For a given likelihood function, VI can be used to obtain an approximated posterior distribution $\upsilon_\phi(z)$ via minimizing $D_{KL}\big[\upsilon_\phi||\mu_\theta\big]$. Suggested by Hoffman (2017), there exists a approximation gap with VI such that MCMC can refine the latent variable z sampled from the approximated posterior $\upsilon_\phi(z)$ to get a higher completed data log-likelihood. We compare the posterior approximations with VI, VI+MCMC (Hoffman, 2017) and SWVI on MNIST data set and report the completed data log-likelihood as evaluation of the approximations. For the likelihood function $p_\theta(x|z)$, we use a decoder from a variational auto-encoder pre-trained on training set. We evaluate the posterior approximations on 100 samples from the test set. For the variational distributions, we use mean-field Gaussian for both VI and SWVI and we also use an implicit variational distribution with neural nets for SWVI. The MCMC method used is Hamiltonian Monte Carlo (HMC) (Neal et al., 2011) More details about the experiment settings can be found in Appendix.

5. Conclusion

We introduced sliced Wasserstein variational inference–a new method of variational inference to minimize a sequence of discrepancies between the variational distribution and the target distribution. SWVI utilizes MCMC to construct such sequence where at each iteration the variational distribution is improved towards the target. We also provide an analysis of the theoretic guarantees to the convergence of the proposed method and justify the consistency when the number of Markov chains goes to infinity. SWVI is flexible where the approximation can be either an implicit black-box sampler or a standard parametric

Method	Complete Data Log-Likelihood
Mean-field VI	-113.78±0.18
Mean-field VI+HMC	-113.53±0.27
Mean-field SWVI	-113.62±0.20
Neural net SWVI	**-113.16**±0.31

Table 2: Comparisons of posterior approximations (10 posterior samples are used to calculate standard deviations). We observe that SWVI with implicit neural net distribution achieves the high log-likelihood.

probability distribution. We illustrate the performance of SWVI on several experiments. SWVI is a general algorithm that applies to density fitting and Bayesian learning problems. In latent generative modeling, SWVI can be used to refine the posterior distributions such that the log likelihood can be improved. In future work, we will study the asymptotic distribution of the estimator obtained via SWVI and its the finite-sample behavior as well. In addition, more practical applications will be further studied such as deep generative modelings and image inpainting.

Acknowledgments

Authors would like to thank all four anonymous reviewers for their insightful comments.

References

Luca Ambrogioni, Umut Güçlü, Yağmur Güçlütürk, Max Hinne, Marcel A van Gerven, and Eric Maris. Wasserstein variational inference. *NeurIPS*, 2018.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017.

Arthur Asuncion and David Newman. UCI machine learning repository, 2007.

Ayanendranath Basu, Hiroyuki Shioya, and Chanseok Park. *Statistical inference: the minimum distance approach*. CRC press, 2011.

Espen Bernton, Pierre E Jacob, Mathieu Gerber, and Christian P Robert. On parameter estimation with the wasserstein distance. *Information and Inference: A Journal of the IMA*, 8(4):657–676, 2019.

Gregory Beylkin. The inversion problem and applications of the generalized radon transform. *Communications on pure and applied mathematics*, 37(5):579–599, 1984.

Niloy Biswas, Pierre E Jacob, and Paul Vanetti. Estimating convergence of markov chains with l-lag couplings. *In NeurIPS*, 2019.

Clément Bonet, Nicolas Courty, François Septier, and Lucas Drumetz. Sliced-wasserstein gradient flows. *arXiv preprint arXiv:2110.10972*, 2021.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51 (1):22–45, 2015.

Nicolas Bonnotte. *Unidimensional and evolution methods for optimal transportation*. PhD thesis, Paris 11, 2013.

Jörg Bornschein and Yoshua Bengio. Reweighted wake-sleep. *arXiv preprint arXiv:1406.2751*, 2014.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, 2013.

Ishan Deshpande, Ziyu Zhang, and Alexander G Schwing. Generative modeling using the sliced wasserstein distance. In *CVPR*, 2018.

Adji Bousso Dieng, Dustin Tran, Rajesh Ranganath, John Paisley, and David Blei. Variational inference via χ^2 upper bound minimization. *In NeurIPS*, 2017.

Yihao Feng, Dilin Wang, and Qiang Liu. Learning to draw samples with amortized stein variational gradient descent. In *UAI*, 2017.

Samuel Gershman and Noah Goodman. Amortized inference in probabilistic reasoning. In *Proceedings of the annual meeting of the cognitive science society*, volume 36, 2014.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. In *NeurIPS*, 2017.

Matthew D Hoffman. Learning deep latent gaussian models with markov chain monte carlo. In *ICML*, 2017.

Ghassen Jerfel, Serena Wang, Clara Fannjiang, Katherine A Heller, Yian Ma, and Michael I Jordan. Variational refinement for importance sampling using the forward kullback-leibler divergence. *arXiv preprint arXiv:2106.15980*, 2021.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

Soheil Kolouri, Phillip E Pope, Charles E Martin, and Gustavo K Rohde. Sliced wasserstein auto-encoders. In *ICLR*, 2018a.

Soheil Kolouri, Gustavo K Rohde, and Heiko Hoffmann. Sliced wasserstein distance for learning gaussian mixture models. In *CVPR*, pages 3427–3436, 2018b.

Yingzhen Li and Richard E Turner. Rényi divergence variational inference. *In NeurIPS*, 29, 2016.

Yingzhen Li, Richard E Turner, and Qiang Liu. Approximate inference with amortised mcmc. *arXiv preprint arXiv:1702.08343*, 2017.

Chang Liu, Jingwei Zhuo, and Jun Zhu. Understanding mcmc dynamics as flows on the wasserstein space. In *ICML*, 2019.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *NeurIPS*, 2016.

Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In *ICML*, 2017.

Kimia Nadjahi, Alain Durmus, Umut Simsekli, and Roland Badeau. Asymptotic guarantees for learning generative models with the sliced-wasserstein distance. *In NeurIPS*, 2019.

Kimia Nadjahi, Alain Durmus, Lénaïc Chizat, Soheil Kolouri, Shahin Shahrampour, and Umut Simsekli. Statistical and topological properties of sliced probability divergences. *In NeurIPS*, 2020.

Christian Naesseth, Fredrik Lindsten, and David Blei. Markovian score climbing: Variational inference with $kl(p||q)$. *In NeurIPS*, 2020.

Radford Neal. *MCMC using Hamiltonian dynamics*. CRC press, 2011.

Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

John Paisley, David Blei, and Michael Jordan. Variational bayesian inference with stochastic search. In *ICML*, 2012.

Dennis Prangle. Distilling importance sampling. *arXiv preprint arXiv:1910.03632*, 2019.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *AISTATS*, 2014.

Rajesh Ranganath, Dustin Tran, Jaan Altosaar, and David Blei. Operator variational inference. *In NeurIPS*, 2016.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *ICML*, 2015.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, 2014.

Francisco Ruiz and Michalis Titsias. A contrastive divergence for combining variational inference and mcmc. In *ICML*, 2019.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

Neng Wan, Dapeng Li, and Naira Hovakimyan. f-divergence variational inference. *In NeurIPS*, 2020.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *ICML*, 2011.

Jacob Wolfowitz. The minimum distance method. *The Annals of Mathematical Statistics*, pages 75–88, 1957.

Quan Zhang, Huangjie Zheng, and Mingyuan Zhou. Mcmc-interactive variational inference. *arXiv preprint arXiv:2010.02029*, 2020.

Appendix A. Proof of Theorem 2

Proof First we introduce lower semi continuity of sliced Wasserstein distance (Lemma 1 of Nadjahi et al. (2019)). For any $\{\mu_t^n\}_{n\in\mathbb{N}}, \{\upsilon_\phi^n\}_{n\in\mathbb{N}} \in \mathcal{P}_p(\mathcal{X})$ weakly converge to μ_t, υ_ϕ, the following inequality holds $\liminf_{n\to\infty} \mathcal{SW}_p(\mu_t^n, \upsilon_\phi^n) \geq \mathcal{SW}_p(\mu_t, \upsilon_\phi)$. Next we introduce **Lemma 4** that gives the sufficient condition that minimum exists.

Lemma 4 *$\mathcal{X}$ is a compact set and $f : \mathcal{X} \to \mathbb{R}$ is lower semi continuous, then f is bounded below and it attains the infimum.*

$B_\epsilon = \{\phi \in \mathcal{H} : \mathcal{SW}_p(\mu_t, \upsilon_\phi) \leq \epsilon^* + \epsilon\}$ is closed since $\phi \to \mathcal{SW}_p(\mu_t, \upsilon_\phi)$ is lower semi-continuous and By **A.4.** we have that B_ϵ is bounded therefore B_ϵ is a compact set. Hence by Lemma 4 we know $\arg\min_{\phi\in\mathcal{H}} \mathcal{SW}_p(\mu_t, \upsilon_\phi)$ is non empty.

Next we prove that $\lim_{n\to\infty} \inf_\phi \mathcal{SW}_p(\mu_t^n, \upsilon_\phi^n) = \inf_\phi \mathcal{SW}_p(\mu_t, \upsilon_\phi)$, the key step is to prove the epi-convergence in sub compact set $\mathcal{K} \in \mathcal{H}$ and open set $\mathcal{O} \in \mathcal{H}$.

$$\begin{cases} \liminf_{n\to\infty} \inf_{\phi\in\mathcal{K}} \mathcal{SW}_p(\mu_t^n, \upsilon_\phi^n) \geq \inf_{\phi\in\mathcal{K}} \mathcal{SW}_p(\mu_t, \upsilon_\phi) \\ \limsup_{n\to\infty} \inf_{\phi\in\mathcal{O}} \mathcal{SW}_p(\mu_t^n, \upsilon_\phi^n) \leq \inf_{\phi\in\mathcal{O}} \mathcal{SW}_p(\mu_t, \upsilon_\phi) \end{cases} \tag{19}$$

For compact set $\mathcal{K}$, by definition such that every sequence in $\mathcal{K}$ has a convergent sub-sequence. By lower semi continuity of $\phi \to \mathcal{SW}_p(\mu_t^n, \upsilon_\phi^n)$, we have $\inf_{\phi\in\mathcal{K}} \mathcal{SW}_p(\mu_t^n, \upsilon_\phi^n) = \mathcal{SW}_p(\mu_t^n, \upsilon_{\phi_n}^n)$ with some $\phi_n \in \mathcal{K}$. Hence

$$\begin{aligned}
&\liminf_{n\to\infty} \inf_{\phi\in\mathcal{K}} \mathcal{SW}_p(\mu_t^n, \upsilon_\phi^n) = \liminf_{n\to\infty} \mathcal{SW}_p(\mu_t^n, \upsilon_{\phi_n}^n) \\
&= \lim_{k\to\infty} \mathcal{SW}_p(\mu_t^{n_k}, \upsilon_{\phi_{n_k}}^{n_k}), \text{ exists sub-sequence converges to } \liminf \\
&= \lim_{l\to\infty} \mathcal{SW}_p(\mu_t^{n_{k_l}}, \upsilon_{\phi_{n_{k_l}}}^{n_{k_l}}) \\
&= \liminf_{l\to\infty} \mathcal{SW}_p(\mu_t^{n_{k_l}}, \upsilon_{\phi_{n_{k_l}}}^{n_{k_l}}) \\
&\geq \liminf_{l\to\infty} \left[\mathcal{SW}_p(\mu_t, \upsilon_{\phi_{n_{k_l}}}^{n_{k_l}}) - \mathcal{SW}_p(\mu_t^{n_{k_l}}, \mu_t)\right], \text{by triangular inequality} \\
&\geq \liminf_{l\to\infty} \mathcal{SW}_p(\mu_t, \upsilon_{\phi_{n_{k_l}}}^{n_{k_l}}) - \limsup_{l\to\infty} \mathcal{SW}_p(\mu_t^{n_{k_l}}, \mu_t) \\
&\geq \mathcal{SW}_p(\mu_t, \upsilon_{\bar{\phi}}), \text{ by Assumption } \textbf{A.3.} \text{ and exists sub-sequence converges in } \mathcal{K} \text{ to } \bar{\phi} \\
&\geq \inf_{\phi\in\mathcal{K}} \mathcal{SW}_p(\mu_t, \upsilon_\phi).
\end{aligned} \tag{20}$$

For open set $\mathcal{O} \in \mathcal{H}$, there exists $\{\phi_n\} \in \mathcal{O}$ such that $\lim_{n\to\infty} \mathcal{SW}_p(\mu_t, \upsilon_{\phi_n}) = \inf_{\phi\in\mathcal{O}} \mathcal{SW}_p(\mu_t, \upsilon_\phi)$. And $\forall n \in \mathbb{N}, \mathcal{SW}_p(\mu_t^n, \upsilon_{\phi_n}) \geq \inf_{\phi\in\mathcal{O}} \mathcal{SW}_p(\mu_t^n, \upsilon_\phi)$. Hence

$$
\begin{aligned}
&\limsup_{n\to\infty} \inf_{\phi\in\mathcal{O}} \mathcal{SW}_p(\mu_t^n, \upsilon_\phi^n) \leq \limsup_{n\to\infty} \mathcal{SW}_p(\mu_t^n, \upsilon_{\phi_n}^n) \\
&\leq \limsup_{n\to\infty} \left[\mathcal{SW}_p(\mu_t^n, \mu_t) + \mathcal{SW}_p(\mu_t, \upsilon_{\phi_n}) + \mathcal{SW}_p(\upsilon_{\phi_n}, \upsilon_{\phi_n}^n)\right], \text{by triangular inequality} \\
&\leq \limsup_{n\to\infty} \mathcal{SW}_p(\mu_t^n, \mu_t) + \limsup_{n\to\infty} \mathcal{SW}_p(\mu_t, \upsilon_{\phi_n}) + \limsup_{n\to\infty} \mathcal{SW}_p(\upsilon_{\phi_n}, \upsilon_{\phi_n}^n), \text{by boundedness} \\
&= \limsup_{n\to\infty} \mathcal{SW}_p(\mu_t, \upsilon_{\phi_n}), \text{ by Assumption } \mathbf{A.3.} = \inf_{\phi\in\mathcal{O}} \mathcal{SW}_p(\mu_t, \upsilon_\phi).
\end{aligned}
$$

Hence we have derived the epi-convergence. It is trivial to prove the convergence under $\mathcal{H}$ $\lim_{n\to\infty} \inf_{\phi\in\mathcal{H}} \mathcal{SW}_p(\mu_t^n, \upsilon_\phi^n) = \inf_{\phi\in\mathcal{H}} \mathcal{SW}_p(\mu_t, \upsilon_\phi)$. We refer to (Bernton et al., 2019; Nadjahi et al., 2019) for similar derivations. ■

Appendix B. Experiment Settings

4.1. Target distributions:

$$
\text{2D Gaussian: } N\left[\begin{pmatrix}1\\2\end{pmatrix}, \begin{pmatrix}1.00 & 0.80\\0.80 & 0.92\end{pmatrix}\right], \text{ Mixture: } 0.5N\left[\begin{pmatrix}-1\\1\end{pmatrix}, \begin{pmatrix}0.80 & 0.00\\0.00 & 0.80\end{pmatrix}\right] + 0.5N\left[\begin{pmatrix}3\\-3\end{pmatrix}, \begin{pmatrix}0.80 & 0.00\\0.00 & 0.80\end{pmatrix}\right]
$$

All variational distributions are initialized at $N\left[\begin{pmatrix}-0.5\\-0.5\end{pmatrix}, \begin{pmatrix}0.50 & 0.00\\0.00 & 0.50\end{pmatrix}\right]$. The number of Markov chains is 300 and starts from the initial variational distribution with lag $L = 20$. The same number of samples are used in VI for reparameterizing gradient estimation.

4.2. Target distributions: $0.5N\left[\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1.00 & 0.00\\0.00 & 1.00\end{pmatrix}\right] + 0.5N\left[\begin{pmatrix}-1\\-2\end{pmatrix}, \begin{pmatrix}0.04 & 0.00\\0.00 & 0.04\end{pmatrix}\right]$. Neural networks have 3 hidden layers (each with 128 units) with input dimension 5. The number of Markov chains is 300 and starts at samples output by the initial Neural networks. Lag $L = 20$. MCMC is random walk Metropolis-Hastings with step size 2.0. SVGD employs 300 particles and uses RBF kernel with empirical median trick bandwidth (Liu and Wang, 2016).

4.4. We pre-train a varitional auto-encoder (VAE) on MNIST using 1,200 randomly selected training samples. The encoder of VAE has one hidden layer with 100 units for both mean and log-variance. The decoder is also one hidden layer neural network with 100 units and with sigmoid activation for the output layer. The latent variable has 10 dimensions.

We fix decoder pre-trained with VAE and set prior distribution to standard Gaussian such that posterior in Eq (18) is defined. 100 randomly selected samples from test set are used for evaluation of posterior approximations. For 'Neural net SWVI', we use a two hidden neural net with units $[128, 64]$ and the input dimension is 5. We parallel 10 HMC chains and HMC takes 3 steps leapfrog with step size 0.1. All HMC chains start at samples generated from the learned mean-field Gaussian via VI. we run HMC 10 times for 'Mean-field VI+HMC', 'Mean-field SWVI' and 100 times for 'Neural net SWVI' (lag $L = 0$ in this experiment).

Remark The number of slices sampled for estimating sliced Wasserstein distance in all experiments is 10 excluding **4.4** where we use 100 slices for 'Neural net SWVI' .